1st Conference on Research Data Infrastructure

Natural Sciences

https://doi.org/10.52825/CoRDI.v1i.283

Published: 07 Sept. 2023

Research Data Management for Experiments in Solid-State Physics: Concepts

Heiko B. Weber[1] [https://orcid.org/0000-0002-6403-9022] , Sandor Brockhauser[2] [https://orcid.org/0000-0002-9700-4803] , Christoph Koch[2] [https://orcid.org/0000-0002-3984-1523] , Laurenz Rettig[3] [https://orcid.org/0000-0002-0725-6696] , Martin Aeschlimann[4] [https://orcid.org/0000-0003-3413-5029] , Walid Hetaba[5] [https://orcid.org/0000-0003-4728-0786] , Marius Grundmann[6] [https://orcid.org/0000-0001-7554-182X] , Markus Kühbach[2] [https://orcid.org/0000-0002-7117-5196] , and Michael Krieger[1] [https://orcid.org/0000-0003-1480-9161]

[1]Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

[2]Department of Physics and IRIS Adlershof, Humboldt-Universität zu Berlin, Germany

[3]Fritz Haber Institute Berlin, Germany [4]RPTU Kaiserslautern-Landau, Germany

[5]Max-Planck-Institute for Chemical Energy Conversion, Mülheim an der Ruhr, Germany

[6]Department of Physics, Universität Leipzig, Germany

Abstract: FAIRmat develops concepts for paving the way to enable FAIR research data in solid-state physics. For selected theoretical data in this field, the NOMAD portal has developed mature concepts and technological solutions for storing data according to the FAIR principles. Extending this approach to experimental data is challenging due to their diversity and missing standards. In this paper we present our comprehensive approach to establish FAIR data in the field of experimental solid-state physics despite its heterogeneity. The concept includes elaboration of standards, community building and methods that facilitate the community's transition to FAIR standards.

Keywords: solid-state physics, research data management, electronic lab notebook, FAIR data

1 Introduction

Within physics, solid-state physics is the largest community. It is extremely heterogeneous, spanning over small groups and large institutes with their professional infrastructure. Technically, a plethora of well-established methods, but also customized, unconventional methods are covered. Still today, metadata are often documented in handwritten laboratory notebooks, and measurement data are predominantly handled in silico, with the data-generating measurement software being often self-written. In contrast, professional stand-alone equipment is predominantly purchased with its own proprietary code, many details of which are intransparent to the user. In summary, the field is technically very advanced, but extremely heterogeneous and diverse, which imposes challenges to modern RDM. In order to enable a transition to FAIR data management in this community, FAIRmat has identified the relevant fields of action, which are presented in this manuscript.

2 Fields of Action

Laboratory Control Software. FAIRmat is currently developing a software (NOMAD CAMELS) that will make it particularly easy for scientists with setting up a new experiment and guarantees FAIR-ready data output. Communication with the instruments and the measurement protocol will be configurable in a graphical user interface. Camel's output is an HDF5 file by default, which contains not only data, but also the technical parameters of the experiment in a structured way.

Electronic Laboratory Notebooks (ELNs). ELNs play a key role in the documentation of experiments, collecting the entity of metadata. In order to enable the the reusability in 'FAIR', ELNs have to become particularly simple to use at two interfaces: First, at the input stage, so that data can be entered accurately and efficiently. Second, at the services which motivate and support users with entering data sufficiently structured and schematised, so that as much automated processing becomes routinely possible. FAIRmat recognises that the community already uses different ELNs and is developing interfaces and tools to structure the data in ELNs. FAIRmat also uses the existing advantages of NOMAD Oasis and offers domain-specific data schemas via NOMAD's ELN customization capabilities, with a clear focus on structured, schematized data handling.

FAIR-ready data management. FAIRmat stresses that data shall always be collected together with metadata as these are the contextualisation which enables a proper interpretation and distilling knowledge (from the data). According to FAIR principles[1], this requires the use of vocabulary which meets community standards and so enables machine interpretability. FAIRmat recognised the requirements of our diverse community, the fast developments of cutting edge experimental techniques. For new methods, FAIRmat suggests and provides tools for a FAIR-ready documentation of all data and metadata. While these documentation may not follow widely accepted community standards, and so the data is not (yet) FAIR by definition, but is digitized and described and may hence be converted to community standards once they are developed.

Community Standards. FAIRmat encourages the different domain scientists to develop community standards required for FAIR data management. FAIRmat contributes to EMglossary harmonisation [2] work organised by Helmholtz Metadata Collaboration, and also proposes experimental-technique-specific domain ontologies for standardisation. These are developed as extensions of the NeXus community standardisation platform[3]. NeXus allows the documentation of data and metadata concepts in a structured and hierarchical way. FAIRmat is also involved in expressing the NeXus standard in the Semantic Web language OWL which supports interoperability by allowing the connection of all data and metadata to other domains or higher-level ontologies. Any experimental data collected and stored according to the NeXus standard can be automatically loaded into the NOMAD RDM solution.

Workflow in the NOMAD environment. NOMAD enables not only a safe and long-term storage of the data, but also to work with the data. FAIRmat provides examples for containerising community software solutions for data reduction, processing and refined analyzes. The containers can be launched in NOMAD to work with the data directly with no need to download or install the domain-specific tools locally but rather access your data at the server. Currently, NOMAD and its local deployment NOMAD Oasis[4] can be installed on a single server or on a Kubernetes[5] cluster to provide the required computing nodes for such remote data analysis work. All uploaded data, surplus those

data generated on the server is indexed by NOMAD and thus is available in searches to those having access rights to the given dataset. After publishing, it becomes publicly accessible and usable under the Create Commons Attribution License (cc-by) 4[6].

Instrument manufacturers as technology partners. A problem for FAIR data is that manufacturers of scientific instruments often provide software that uses proprietary formats. It is very tedious and sometimes impossible to link data and necessary metadata with the entries in the ELNs and to generate FAIR data out of proprietary sources. FAIRmat takes the approach of developing application-specific data schemes within the scientific community and presenting them to the technology partners. In joint workshops with the companies, data experts and scientists, the details are then discussed and a canon is defined for how the data should be stored in the future.

Broad data expertise. Neither FAIRmat nor local data stewards can solve the plethora of tasks that arise when our community converts all processes to FAIR data. This calls attention to the scientists themselves. In physics education, data handling is not yet considered a crucial skill. Here, a discussion was initiated within the physics community to consider data competence as an additional key competency. For example, on the initiative of FAIRmat scientists, the use of ELNs was introduced in bachelor lab courses[7]. This leads to enhanced competences of the students, and establishes ELNs as the new working standard. Moreover, lab course ELNs are a convenient sandbox in which new RDM concepts and technologies can be tried out within a time-limited framework. Only with the active cooperation of all scientists involved a sustainable transition to FAIR data can succeed.

Funding

FAIRmat is funded by the the Deutsche Forschungsgemeinschaft "DFG, German Research Foundation" – project 460197019.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. DOI: 10.1038/sdata.2016.18.

[2] *Em glossary initiative*, 2023. [Online]. Available: https://helmholtz-metadaten.de/en/em-glossary-initiative.

[3] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, *et al.*, "The nexus data format," *Journal of applied crystallography*, vol. 48, no. 1, pp. 301–305, 2015. DOI: 10.1107/S1600576714027575.

[4] *Nomad - materials science data managed and shared*, 2023. [Online]. Available: https://nomad-lab.eu/.

[5] *Kubernetes, also known as k8s, is an open-source system for automating deployment, scaling, and management of containerized applications.* 2023. [Online]. Available: https://kubernetes.io/.

[6] *Creative commons attribution 4.0 international*, 2023. [Online]. Available: https://creativecommons.org/licenses/by/4.0/.

[7] M. Krieger, H. B. Weber, and C. van Eldik, "Früh zur Datenkompetenz," *Physik Journal*, vol. 21, p. 42, 2022. [Online]. Available: https://www.pro-physik.de/restricted-files/158142.